**Học máy thống kê (DS102)**

THỰC HÀNH TUẦN 3

PHÂN LỚP VĂN BẢN CƠ BẢN

Mssv: 20522090

Name: Phạm Lê Trưởng

**Chart, bar chart

Description automatically generatedLabels:**

**0: Positive , samples=159**

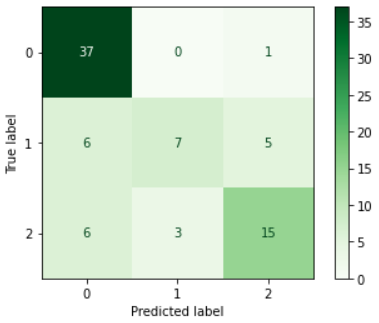
**1: Neutral , samples=98**

**2: Negative , samples=143**

**Naive Bayes model with CountVectorizer (1-gram)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1** |
| 0 - Positive | 0.76 | 0.97 | 0.85 |
| 1 - Neutral | 0.70 | 0.39 | 0.50 |
| 2 - Negative | 0.71 | 0.62 | 0.67 |
| Macro avg | **0.72** | **0.66** | **0.67** |

Accuracy = 0.75

****

**Naive Bayes model with CountVectorizer (2-gram)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1** |
| 0 - Positive | 0.60 | 0.87 | 0.71 |
| 1 - Neutral | 0.44 | 0.22 | 0.30 |
| 2 - Negative | 0.69 | 0.46 | 0.55 |
| Macro avg | **0.58** | **0.52** | **0.52** |

Accuracy = 0.60

Chart

Description automatically generated

**Naive Bayes model with TfIdfVectorizer (1-gram)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1** |
| 0 - Positive | 0.74 | 0.97 | 0.84 |
| 1 - Neutral | 1.00 | 0.28 | 0.43 |
| 2 - Negative | 0.72 | 0.75 | 0.73 |
| Macro avg | **0.82** | **0.67** | **0.67** |

Accuracy = 0.75

Chart

Description automatically generated

**Naive Bayes model with TfIdfVectorizer (2-gram)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1** |
| 0 - Positive | 0.59 | 0.89 | 0.71 |
| 1 - Neutral | 0.67 | 0.11 | 0.19 |
| 2 - Negative | 0.58 | 0.46 | 0.51 |
| Macro avg | **0.61** | **0.49** | **0.47** |

Accuracy = 0.59

Chart

Description automatically generated

Kết quả model **TfIdfVectorizer (1-gram)** tốt nhất

**Accuracy = 0.75**

**Precision = 0.82**

**Recall = 0.67**

**Nhận xét:**

**TfIdfVectorizer** tốt hơn so với **CountVectorizer** vì nó không chỉ tập trung vào tần suất của các từ có trong vocabulary mà còn cung cấp tầm quan trọng của các từ. Sau đó có thể loại bỏ các từ ít quan trọng hơn để phân tích, do đó làm cho việc xây dựng mô hình ít phức tạp hơn bằng cách giảm kích thước đầu vào.